The UMLS® as a Domain Model of Medicine

Thomas C. Rindflesch, PhD
Alan R. Aronson, PhD
William T. Hole, MD
National Library of Medicine, Bethesda, MD 20894

The UMLS Metathesaurus contains a vast amount of information which can potentially serve as the basis for programs designed to provide improved access to online biomedical information. We investigate the structure of this information and suggest some natural language processing techniques for addressing the complexity of Metathesaurus concepts. The methods discussed assign a semantic interpretation to complex Metathesaurus concepts and rely on the UMLS Semantic Network as well as the Metathesaurus itself.

1. INTRODUCTION

Our goal is to explore some of the issues involved in using the UMLS [1] as a domain model of medicine. There is a widespread point of view that extensive, structured domain knowledge is essential for significant advances in automatic processing of free text [2,3]. A domain model satisfies this need by defining the entities and relationships in some world [4,5]. In an ideal domain model breadth of coverage is extensive, relationships are encoded as explicit rules, and entities are atomic concepts or, for complex concepts, internal relationships are explicitly defined. In the medical domain, the GALEN project [6,7] is under development, while [8] discusses the use of the UMLS as a medical knowledge source. In this paper we concentrate on an approach to dealing with complex concepts in the UMLS by identifying atomic concepts and the relationships among them.

The 1998 version of the UMLS Metathesaurus® is a compendium of over 40 controlled vocabularies in the biomedical domain containing a vast amount of information, pertinent to a wide variety of interests. For example, there are 66,451 English strings representing concepts having semantic type ‘Therapeutic or Preventive Procedure’ (see [9] for clinically pertinent terms). The information represented in the Metathesaurus is structured in various ways, partly by explicit information added by the editors, and partly by explicit information inherent in the constituent vocabularies. The structured information in the Metathesaurus is enhanced by the UMLS Semantic Network, which stipulates various relationships between classes of Metathesaurus concepts through the semantic types assigned to each concept. There is also a great deal of implicit structural information, based on the individual vocabularies.

In a sense, there is a domain model of medicine “hidden” in the UMLS Metathesaurus and Semantic Network, taken together. The relationship between the domain model and the UMLS is not quite one-to-one because not all Metathesaurus concepts are atomic, and further not all relationships expressed in complex concepts are explicit. After a brief consideration of complexity below, we propose a general method for discovering and explicitly marking the domain model of medicine in the UMLS.

We assume that the only criterion for complexity is syntactic, and that there is thus a regular relationship between syntactic complexity and conceptual complexity. In particular our work is based on the principle that entries in a lexicon are conceptually simple (or atomic). We realize that this may not actually be so in all instances, but we feel it is a reasonable operational assumption. It then follows that a syntactic structure which consists of a single lexical entry is simple and any structure composed of more than one lexical entry is complex. It is further the case that all com-

plex structure fall into two classes: those that exhibit canonical English syntactic structure and those that do not.

On the basis of the preceding considerations, our strategy for analyzing structural complexity is to use the explicit information in the Metathesaurus to partition all English strings into two groups: those which are likely to display canonical English syntactic structure and those which are not. We then apply natural language processing techniques to the first group in an attempt to identify atomic concepts and to make explicit the implicit relationships exhibited by strings representing complex concepts (see [10-12] for related approaches). Although we do not actively address recovering the implicit relationships expressed in strings with noncanonical English structure, it is clear that the first step in attempting such a task is the identification of these strings.

2. SPECIALIST™ NATURAL LANGUAGE PROCESSING

The SPECIALIST system [13] provides a framework for research aimed at exploiting the resources of the UMLS in processing biomedical text. In addition to the Metathesaurus and Semantic Network, the SPECIALIST Lexicon and associated lexical variant programs [14] as well as the Knowledge Source Server [15] support syntactic analysis and semantic interpretation of free text in the biomedical domain.

The SPECIALIST system begins analysis of biomedical text by consulting the Lexicon to determine syntactic information for each lexical entry in the input. This information is then given to a stochastic tagger [16] for resolution of part-of-speech ambiguities. An underspecified syntactic analysis [17] is then produced on the basis of the lexical information (with most ambiguities resolved by the tagger). For example, input text *pancreatic secretory trypsin inhibitor* is given the following analysis: [[mod(pancreatic), mod(secretory), mod(trypsin), head(inhibitor)]]

In particular, note that, although the head of the noun phrase and its modifiers have been identified, no indication is given of the internal syntactic structure of such phrases. It is our experience, that this attenuated analysis is sufficient to serve as the basis for usable semantic interpretation.

The next step in processing text calls MetaMap [18], a program for mapping free text to concepts in the Metathesaurus. This program takes advantage of syntactic processing and considers each noun phrase individually as it proceeds. For example, it takes as input the underspecified syntactic analysis of *ligation of aorta* and finds the following Metathesaurus concepts:

(1) Ligation
('Therapeutic or Preventive Procedure')
Aorta
('Body Part, Organ, or Organ Component')

The program SemRep then performs semantic processing [17]. It depends on both syntactic analysis and the Metathesaurus concepts provided by MetaMap. In addition, it consults the Semantic Network as part of the process of producing a final semantic interpretation. For example, in assigning an interpretation to *ligation of aorta*, the semantic interpreter notes the syntactic analysis given for this input and then consults a rule which states that the preposition *of* corresponds to the Semantic Network relation LOCATION_OF, and further notes that one of the relationships in the Semantic Network with this predicate is:

(2)

Semantic Type 1:	'Body Part, Organ, or Organ Component'
Relation:	LOCATION_OF
Semantic Type 2:	'Therapeutic or Preventive Procedure'

The MetaMap output for this input is then referred to and it is noted that the semantic type for the Metathesaurus concept for the text phrase *ligation* is 'Therapeutic or Preventive Procedure' and that the type for the phrase *aorta* is 'Body Part, Organ, or Organ Component.' Since these semantic types match those found in the relationship indicated by the preposition *of* (LOCATION_OF), (3) is produced as the semantic interpretation for this phrase, where the corresponding Metathesaurus concepts are substituted for the semantic types in the Semantic Network relationship.

(3) Aorta-LOCATION_OF-Ligation

All of the output produced by the preceding linguistic analysis when applied to Metathesaurus strings is exploited during the process of identifying atomic concepts and attempting to discover the internal structure of complex concepts.

3. FILTERING BASED ON STRING TYPE AND LEXICAL CONSIDERATIONS

We began the search for a domain model in the UMLS by considering the 923,841 English strings in the 1998 version of the Metathesaurus. Initial filtering, prior to linguistic processing, eliminates a large number of strings based on lexical considerations and on Metathesaurus-assigned string type values. In "lexical filtering," two strings which differ only by one of the following criteria are normalized to one of the variants: case variation, NEC/NOS variation, uninversion, hyphen variation, non-essential parentheses.

Further filtering takes advantage of explicit information about strings in the Metathesaurus. 94,095 strings have Term Status (TS) of lowercase *s*. These "suppressible synonyms" were eliminated from further processing. An additional 31,002 strings having various values of Term Type were also dropped, including those having the following values: AB (Abbreviation in any source vocabulary); LN (LOINC official fully specified name) and LX (Official fully specified name with expanded abbreviations); and OA (Obsolete abbreviation). After lexical and type filtering 661,768 strings remained. Of these, 174,270 were MeSH® Supplemental Chemicals which we put into a separate partition for later processing. We thus began linguistic processing on 487,498 Metathesaurus strings.

4. LINGUISTIC ANALYSIS OF METATHESAURUS STRINGS

A first pass at categorizing Metathesaurus strings according to their complexity is based on syntactic structure. As determinants of complexity, we refer to number of phrases and number of syntactic items, where phrases are the number of constituents determined by the underspecified analysis and syntactic items are the components in these phrases. As an example, the string *aneurysm of artery of upper extremity,* which is given the syntactic analysis in (4) has three syntactic (noun) phrases and five syntactic items: head, prep, head, prep, and head.

(4) [[head(aneurysm)], [prep(of), head(artery)], [prep(of), head(upper extremity)]]

Based on these criteria the 487,498 Metathesaurus strings which we submitted to linguistic analysis can be categorized as follows, and the succeeding discussion considers each of these classes in turn:

- 292,794 strings can be analyzed into a single syntactic phrase
  - 65,387 contain a single syntactic item

- 119,342 contain exactly two syntactic items
- 108,065 contain more than two syntactic items

• 194,704 strings were analyzed as having more than one syntactic phrase

4.1 A single phrase with a single syntactic item

292,794 strings can be analyzed into a single simple syntactic phrase (all are noun phrases). The simplest subtype of the single simple noun phrase is characterized by those strings analyzable as containing a single syntactic item (65,387), such as *thorax,* or *glomerulosclerosis.* For the majority of these the single syntactic item is head (53,275). This does not mean, however, that these phrases necessarily contain a single word. There are 22,118 multiple-word lexical items (such as *biological oscillators*, or *high blood pressure)* functioning as heads of Metathesaurus strings analyzable as simple noun phrases (31,157 contain single-word heads). Such "lexicalized phrases" are listed in the SPECIALIST Lexicon.

4.2 A single phrase and exactly two syntactic items

119,342 strings can be analyzed as a single syntactic phrase containing exactly two syntactic items. The vast majority of these represent actual syntactic patterns and can be further processed with NLP techniques to identify the constituent simple concepts. The crucial characteristic of these strings in determining whether they are amenable to further processing is position of the head: In almost 95,000 strings with two syntactic items the head is final, the more normal position. These can readily be subjected to further analysis, some of which is discussed below.
As noted, most of the strings analyzed as containing two syntactic items contain a head canonically in final position. Of these, the majority (81,432) have a head preceded by a single modifier. An experiment was conducted on a subset of the strings containing a modifier preceding the head to determine their composition. There are 115 such strings where the head is the word *cancer*, such as *thyroid cancer.*
These strings were subjected to further semantic processing to determine whether the syntactic items (modifier and head) could be mapped to simple concepts in the Metathesaurus, and further, whether the Semantic Network stipulated a relationship between these constituents. In the vast majority of these strings, the modifying concept was found separately in the Metathesaurus, although a few strings like *precancerous* do not occur separately.
After consulting the Semantic Network, SemRep determined the appropriate semantic relationship between the constituent concepts for 100 of the 115 strings containing *cancer* as the head. In 92 instances the (correct) relationship found was LOCATION_OF, as seen in the interpretations below for the strings given above.

(5) Thyroid Gland-LOCATION_OF-Cancer <1>

Seven of the phrases processed stipulate that cancer is an issue in the specialty expressed in the modifier, and the interpretation reflects this assertion, as seen in the following interpretation for *pediatric cancer.*

(6) Cancer <1>-ISSUE_IN-Pediatrics

Certain terms did not receive a semantic interpretation because the modifying concept, although found in the Metathesaurus, was not in a relationship with *cancer* defined in the Semantic Network. Examples are *green cancer, generalized cancer,* and *multiple cancer.*

4.3 A single phrase with more than two syntactic items

108,065 strings with one syntactic phrase have more than two syntactic items. These strings display a variety of patterns, some of them linguistic and some not. The non-linguistic patterns can often be segregated using a combination of source vocabulary, semantic type, and patterns of syntactic items. For example, strings containing punctuation and having semantic type 'Amino Acid, Peptide, or Protein' from LOINC form a natural class which could be processed further: *carnitine.free, carnitine.total.* Similarly, a sizeable number of terms from RCD95 have semantic type 'Immunologic Factor' and display a regular pattern: *a*0101, a*0201, a*0202*.
39,863 strings consisting of one noun phrase have three syntactic items and those items are modifier, modifier, head. Depending on semantic type these can be further processed with linguistic tools. For example, of these, 7,347 have semantic type 'Disease or Syndrome' and in 329 of these the first modifier is the word *acute*. In order to determine how many of these strings were complex, and if they were what their internal structure was, *acute* was removed from these strings and the resulting substring (which now had the structure [modifier, head]) was searched in the Metathesaurus. 172 such strings were found and were not reprocessed. 157 such strings were not found in the Metathesaurus, and these were reprocessed. Many of these followed the pattern seen in strings beginning with *bacterial,* for example: *bacterial bronchitis*. For all of these strings, the constituents occur in the Metathesaurus and the relationship between the two stipulated by the Semantic Network was determined by semantic processing:

(7) Bacteria <1>-CAUSES-Bronchitis

4.4 More than one syntactic phrase

194,704 Metathesaurus strings can be analyzed as having more than one syntactic phrase. These strings are often quite long and present a particular challenge to further processing. However, certain profitable tactics are feasible, based first on explicit information either in the Metathesaurus or encoded from the constituent vocabularies. For example, of the 194,704 strings with more than one syntactic unit, 13,921 represent concepts having one of the semantic types which are children of 'Substance' (excluding 'Body Substance') in the Semantic Network. These are often chemical names requiring special consideration, and we have partitioned these for separate processing. Another, well-defined class of strings in this group is indicated by those representing concepts having semantic type 'Medical Device' (6,984 such strings). Many of these can be analyzed as normal English syntactic structures, such as *conduit with homograft valve.* For those which do not represent canonical English syntax, when the specific vocabularies are consulted, a number of generalizations appear which render many strings tractable to further processing.
For the strings containing more than one phrase which do represent normal English syntax, effective linguistic processing is often possible. For example, of the strings having more than one syntactic phrase, 2,690 have the following characteristics: The first phrase is a simple noun phrase consisting solely of a head ([head]) and the second is a simple prepositional phrase ([preposition,head]). Further, the string represents a concept having semantic type 'Therapeutic or Preventive Procedure' from SNOMED. These strings were subjected to semantic processing and 1,569 received an interpretation. An example is given in (8), which includes the original string, constituent Metathesaurus concepts, and semantic interpretation.

(8) a. suture of bile duct

b. Suture, NOS
Sutures
Bile Ducts

c. Bile Ducts-LOCATION_OF-Suture, NOS

A similar experiment was run on another group of strings from SNOMED representing concepts with semantic type 'Therapeutic or Preventive Procedure.' These consisted of two phrases with the second phrase being a simple prepositional phrase, as above. However, the first phrase was slightly more complicated in that it contained a modifier in addition to a head. There are 707 such phrases, such as *segmental osteoplasty of maxilla* and *orthopedic procedure on head.* Many of these received at least a partial semantic interpretation, and at least some were completely interpreted as in (9).

(9) a. percutaneous balloon valvuloplasty of aortic valve

b. Aortic Valve-LOCATION_OF-Percutaneous valvuloplasty
Percutaneous valvuloplasty-USES-Balloon

Many of the long strings representing complex syntactic structure are not amenable to complete linguistic interpretation. However, partial processing can begin to control the complexity. As an initial inspection of the potential in this regard we extracted all the strings with more than one syntactic phrase representing CPT concepts having semantic type 'Therapeutic or Preventive Procedure' (1,523 strings). Although many of these have a straightforward syntactic structure, (10a) represents the degree of complexity of a fair number of these strings. Note that although the semantic interpretation in (10c) is far from complete, it does hint at the assertion expressed by the string. Furthermore, all of the simple constituent concepts have been found by MetaMap (10b).

(10) a. surgical operation with transplant of whole organ causing abnormal patient reaction, or later complication, without mention of misadventure at time of operation

b. Surgical
Operative Procedures <1>
Transplant, NOS
organ
Abnormal
Patients
Late
Complication, NOS
complications <1>
Time
Operative Procedures <1>

c. Operative Procedures <1>-ISSUE_IN-Surgical
Transplant, NOS-PART_OF-organ

5. CONCLUSION

The preceding analysis appears to indicate the feasibility of determining, for a sizeable number of Metathesaurus strings, whether a concept is atomic or amenable to internal analysis. The 65,387 strings analyzed as a single syntactic item in a single syntactic phrase, whether or not found in the

Lexicon, can be considered atomic concepts in the model. There is some indeterminacy as to whether all lexicalized phrases should be left unanalyzed, although it is probably safest to do so. Determining the internal relations in such concepts may not be reliable with current automatic processing techniques. For strings with complex (but canonical) syntactic structure, semantic processing of the type discussed here provides a valuable tool on which to base further processing. Such processing could relate many of the concepts represented by complex strings to simple Metathesaurus concepts by a principle which states that if the first head of a complex concept matches the head of a simple concept (and the two concepts have the same semantic type), then the complex concept is related to the simple one in the relationship specified by the semantic interpretation of the complex concept. That is, "Suture of arteriovenous fistula" is related to "Sutures, NOS" by the LOCATION_OF relationship (which is a "narrower than" relationship). Similarly, "Skin cancer" is related to "Cancer" also by the LOCATION_OF relationship. A related principle states that if two complex structures have the same semantic interpretation, they are synonyms. For example, *enucleation of eye* (Eye-LOCATION_OF-Enucleation) and *eye enucleation* (Eye-LOCATION_OF-Enucleation).

Suggestions for at least categorizing other classes of Metathesaurus strings with regard to inclusion in the model have been made in two instances. Substances (other than body substances) fall into a natural class which can be isolated and subjected to further processing with consequences for the domain model. A further class of strings, namely some names for medical devices in certain vocabularies can also be segregated prior to human review with regard to the model. We feel that the research we are pursuing can serve as the basis for making the vast amount of knowledge contained in the UMLS Metathesaurus more easily accessible to advanced programs designed to provide enhanced access to online biomedical information.

References

1. Humphreys BL, Lindberg DAB, Schoolman HM, and Barnett GO. The Unified Medical language System: An informatics research collaboration. *JAMIA* 5:1, 1998, 1-13.
2. Bates M and Weischedel RM. *Challenges in natural language processing.* Cambridge: Cambridge University Press, 1993.
3. Rassinoux AM, Wagner JC, Lovis C, Baud RH, Rector A, and Scherrer JR. Analysis of medical texts on a sound medical model. In Gardner RM (ed.) *Proceedings of the 19th Annual SCAMC,* 1995:27-31.
4. Quillian MR. Word Concepts: A Theory and Simulation of Some Basic Semantic Capabilities. In Brachman RJ and Levesque HJ (eds) *Readings in Knowledge Representation*. Los Altos, CA: Morgan Kaufmann, 1985, 98-118.
5. Sowa JF. *Conceptual Structures: Information Processing in Minds and Machines*. Reading, MA: Addison-Wesley, 1984.
6. Rector A, Salomon W, Nowlan W, Rush T, Zanstra P, and Classen W. A medical terminology server for medical language and medical information systems. *Methods of Information in Medicine* 1995;34(1):147-157.
7. Rogers JE and Rector AL. Terminological Systems: Bridging the Generation Gap. In Masys DR (ed.) *Proceedings of the 1997 AMIA Annual Fall Symposium,* 1997, 610-614.

8. Bodenreider O, Burgun A, Botti G, Fieschi M, LeBeux P, and Kohler F. Evaluation of the Unified Medical language System as a medical knowledge source. *JAMIA* 5:1, 1998, 76-87

9. Humphreys BL, McCray AT, and Cheh ML. Evaluating the Coverage of Controlled Health Data Terminologies: Report on the Results of the NLM/AHCPR Large Scale Vocabulary Test. *JAMIA* 4:6, 1997, 484-500.

10. Riloff E. An Empirical Study of Automated Dictionary Construction for Information Extraction in Three Domains. *Artificial Intelligence* 85, 1996, 101-134.

11. Cimino JJ. Auditing the Unified Medical Language System with semantic methods. *JAMIA* 5:1, 1998, 41-51.

12. Dolin RH, Huff SM, Rocha RA, Spackman KA, and Campbell KE. Evaluation of a "Lexically Assign, Logically Refine" Strategy for Semi-automated Integration of Overlapping Terminologies. *JAMIA* 5:2, 1998, 203-213.

13. McCray AT, Aronson AR, Browne AC, Rindflesch TC, Razi A and Srinivasan S. UMLS knowledge for biomedical language processing. *Bulletin of the Medical Library Association* 81, 1993, 184-194.

14. McCray AT, Srinivasan S and Browne AC. Lexical methods for managing variation in biomedical terminologies. In Ozbolt JG (ed.) *Proceedings of the 18th Annual SCAMC*, 1994, 235-239.

15. McCray AT, Razi AM, Bangalore AK, Browne AC, and Stavri PZ. The UMLS Knowledge Source Server: A Versatile Internet-Based Research Tool. In Cimino JJ (ed.) *Proceedings of the 1996 AMIA Annual Fall Symposium*, 1996, 164-168.

16. Cutting D, Kupiec J, Pedersen J and Sibun P. A practical part-of-speech tagger. In *Proceedings of the Third Conference on Applied Natural Language Processing*, 1992.

17. Rindflesch TC and Aronson AR. Semantic processing in information retrieval. In Safran C (ed.) *Proceedings of the 17th Annual SCAMC*, 1993, 611-615.

18. Aronson AR, Rindflesch TC, and Browne AC. Exploiting a large thesaurus for information retrieval. *Proceedings of RIAO 94,* 1994:197-216.